面向低资源语言机器翻译的平行语料句对齐评分

李林霞[1，3]，陈波[2，3]，周毛克[1，3]，赵小兵[2，3]
[1](中央民族大学中国少数民族语言文学学院 北京 100081)
[2](中央民族大学信息工程学院 北京 100081)
[3](中央民族大学国家语言资源监测与研究少数民族语言中心 北京 100081)

摘要:

[目的]量化低资源语言平行语料的句对齐评分，获取高质量平行语料，提升机器翻译的性能。

[方法]提出基于神经网络的无监督句嵌入双语平行语料句对齐评分方法 NeuroAlign：将平行句对嵌入至同一向量空间，计算平行语料中给定候选句对的对齐评分，然后根据评分排序过滤分值较低的平行句对，获得高质量的低资源语言双语平行语料。

[结果]BUCC2018 平行文本挖掘任务中 F1 值可提升 0.5-0.8；CCMT2021 低资源语言神经机器翻译中 BLEU 值可提升 0.1-10.9；句对齐评分可接近人工评分。

[局限]限于低资源双语平行语料的资源匮乏，未在藏汉、维汉、蒙汉以外的语言对上进行探索研究。

[结论]可以有效应用至低资源语言平行语料的句对齐评分，从数据源端提升语料质量，进而改进机器翻译的效果。

关键词: 机器翻译；低资源语言；平行语料；句对齐评分

分类号: TP393，G250

DOI: 10.11925/infotech.2096-3467.2024.0065

Parallel Corpus Sentence Alignment Scoring for Low-Resource Language Machine Translation

Li Linxia[1,3], Chen Bo[2,3], Zhou Maoke[1,3], Zhao Xiaobing[2,3]
[1](School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, China)
[2] (School of Information Engineering, Minzu University of China, Beijing 100081, China)
[3] (National Language Resource Monitoring and Research Center of Minority Languages, Minzu University of China, Beijing 100081, China)

Abstract:

[Objective]This paper aims to quantify the sentence alignment scores of low-resource parallel corpora to obtain high-quality parallel corpora, improving machine translation performance.

[Methods] We propose NeuroAlign, a neural network-based unsupervised sentence embedding method for scoring bilingual parallel sentence alignment. Parallel sentence pairs are embedded into the same vector space, and alignment scores for given candidate sentence pairs in the parallel corpus are calculated. Based on these scores, low-scoring sentence pairs are filtered out, resulting in high-quality bilingual parallel corpora for low-resource languages.

[Results] In the BUCC2018 parallel text mining task, the F1 score can be improved by 0.5-0.8. In the CCMT2021 low-resource language neural machine translation task, the BLEU score can be improved by 0.1-10.9. The sentence alignment scores can approach human evaluation.
[Limitations] Due to the scarcity of low-resource bilingual parallel corpora, research has not been conducted on language pairs other than Tibetan-Chinese, Uyghur-Chinese, and Mongolian-Chinese.
[Conclusions] This method can be effectively applied to sentence alignment scoring for low-resource language machine translation parallel corpora, improving the quality of the data source, and thereby enhancing machine translation performance.
Keywords： Machine Translation; Low-Resource Language; Parallel Corpus; Sentence Alignment Scoring

1 引言

机器翻译系统训练需要大量语义相同的双语平行文本语料（简称平行语料），通常认为平行语料规模越大机器翻译的效果越好。事实上，除了数据规模外，平行语料的其他质量因素也会对机器翻译性能产生影响[1]，如：领域分布、句对齐质量等。研究表明，统计机器翻译（Statistic Machine Translation, SMT）中平行语料的对齐错误会影响系统性能[2]；神经机器翻译（Neural Machine Translation，NMT）中平行语料之间的未翻译和错位对系统性能影响更大[3]。因此，平行语料的句对齐质量是影响机器翻译系统性能的重要因素之一。

早期基于特征工程的平行语料句对齐技术取得了一定成功，但这种方法相对繁琐，且捕捉到的特征不一定准确，从而会影响后期的翻译效果。随着 NMT 的发展，句对齐任务受到的关注虽不如 SMT 以及一些监督和半监督 NMT 方法高，但并不意味着平行语料的句对齐质量对机器翻译的影响减小了。在当前流行的“预训练+微调”[4-5]研究范式下，尽管大规模预训练语言模型在机器翻译任务上表现出色，看似减少了对平行语料的依赖。但是，目前除了中文、英文等资源丰富的语言外，世界上绝大多数语言都缺乏大规模、高质量的平行语料，大部分语言仍存在平行语料稀缺的现实困境，即便是现有的一些低资源语言可以通过少量的平行语料微调训练，也要保证这些平行语料的数据质量。

通过对 CCMT2021[1]低资源语言平行语料分析发现，即便是官方公布的质量较好的平行语料，也不免存在未对齐、未翻译、语言错误、翻译错误、断句错误、编码错误等问题。因此，本文受平行语料挖掘任务的启发，针对上述平行语料中出现的未对齐问题，提出了一种基于神经网络的无监督句嵌入双语平行语料句对齐评分方法[6-8]NeuroAlign（**Neural** Network-based Sentence Embedding **Alignment** Scoring Method for Bilingual Parallel Corpora, 简称 NeuroAlign），旨在评价双语平行语料的句对齐质量，期望通过评分为低资源语言机器翻译进行语料筛选。

该方法首先将双语平行语料嵌入至同一向量空间，通过计算源语言（Source Language，S）和目标语言（Target Language，T）给定候选句对的余弦相似度（给定候选余弦），与其最邻近的 k 个候选余弦（邻近候选余弦）之间的比率差值来判断两个句子的对齐程度。同时，对平行句对进行了句长惩罚，使过短或过长的句子在计算句对齐评分时不会占据过大的优势或劣势。实验证明，该方法在高资源语言平行语料挖掘、低资源语言神经机器翻译和句对齐评分三个任务上均表现

1 第十七届全国机器翻译大会（The 17th China Conference on Machine Translation, CCMT2021）

出较高性能，且该方法的应用不需要修改翻译系统的模型框架，只需从数据源端筛选高质量的平行语料就可提升机器翻译性能。

2 相关工作

2.1 平行语料对齐

早期的平行语料获取方式主要通过高度工程化的系统[9]收集，后期的方法则侧重于文本内容，可以通过基于知识的语料收集[10]、平行语料挖掘[6]等方法来获取，其中句对齐方法属平行语料挖掘的方法之一。

句对齐任务是指从平行语料中识别一个句子与另一种语言句子之间的对应关系。通常会先定义一个对齐的评分函数，然后使用动态规划算法[11]最大化全局对齐分数，输入是一对文本，输出是句子之间的假设对齐方式。早期的对齐方法主要基于统计特征信息，如：句长[12-13]、词汇[14-15]或部分信息对齐[16]。基于词汇的方法容易受到语言的限制，不同的语言需要提取不同的特征；基于句长的对齐方法在句子长度相同的情况下表现不佳。随着深度学习发展，出现了基于神经网络的对齐方法[17-20]，并取得了很好的效果，其核心思想是利用嵌入空间向量的相似度来进行对齐[8、21-23]，但具体的嵌入方法和对齐算法各有不同。

Melvin Johnson 等[24]采用多语言句嵌入来编码多种语言，训练多语言机器翻译（包含一对多、多对一、多对多），编码时需要在每一种语言前加入一个语言标签来区分不同语言。Schwenk[8]没有使用特殊的输入标记来指示不同的目标语言，而是通过共享编码器学习了联合多语言句嵌入，将 9 种语言的完整句子嵌入到联合空间，并使用不同语言句子之间的距离阈值来过滤和挖掘不同语言对之间的平行语料。与之前方法不同的是，本文侧重于用句对齐方法来实现对双语平行语料的句对齐量化评分，期望通过评分的方式为低资源语言机器翻译进行语料筛选，从数据源端提升机器翻译的效果。

2.2 句对齐评分指标

平行语料的句对齐评分可以采用人工评估或自动评估两种方式，其中自动评估分为以下两种情况：

第一种，已知有 n 对相互翻译的词、句、篇章时，平行语料的对齐评分通常以准确率（Precision）、召回率（Recall）、F1 值（F1-score）来判断效果。

第二种，无法确定翻译的词、句、篇章数量时，采用一些间接的指标来给定对齐评分，如：译文评价指标（如：Bilingual Evaluation Understudy, BLEU[25]）、余弦相似度等。

Moore[14]、Varga[15]等采用了翻译对齐的思想，将两个文档转换为同一语言，并引入修正后的机器翻译译文评价指标 BLEU 来判断文本对齐效果或挖掘出平行语料[5-6、19-20]。

采用句嵌入的方法通常用余弦相似度结合固定阈值来衡量句子之间的对齐程度。Artetxe 和 Schwenk[6]认为这种对齐方法会产生余弦相似度得分范围不一致的问题，即对齐错误的句子比对齐正确的句子具有更大的余弦相似度，从而不利于用固定阈值来过滤句子，因此研究者们提出了不同的基于余弦相似度的修正评分[6-7]。本文在 Artetxe 和 Schwenk[6]提出的算法基础上，分别对源语言和目标语言句向量进行了平滑，在向量空间中拉近了给定候选余弦与邻近候选余弦之间的比率差值，从而对句对齐质量进行了更为严格的评分。

3　研究内容

本文提出了基于神经网络的无监督句嵌入双语平行语料句对齐评分方法。首先，采用神经网络的句嵌入方法，将平行句对嵌入至同一向量空间，然后，计算平行语料的句对齐评分，再根据评分排序过滤对齐评分较低的平行句对，以此来获取相对高质量的低资源语言双语平行语料数据集。具体流程如图 1 所示。

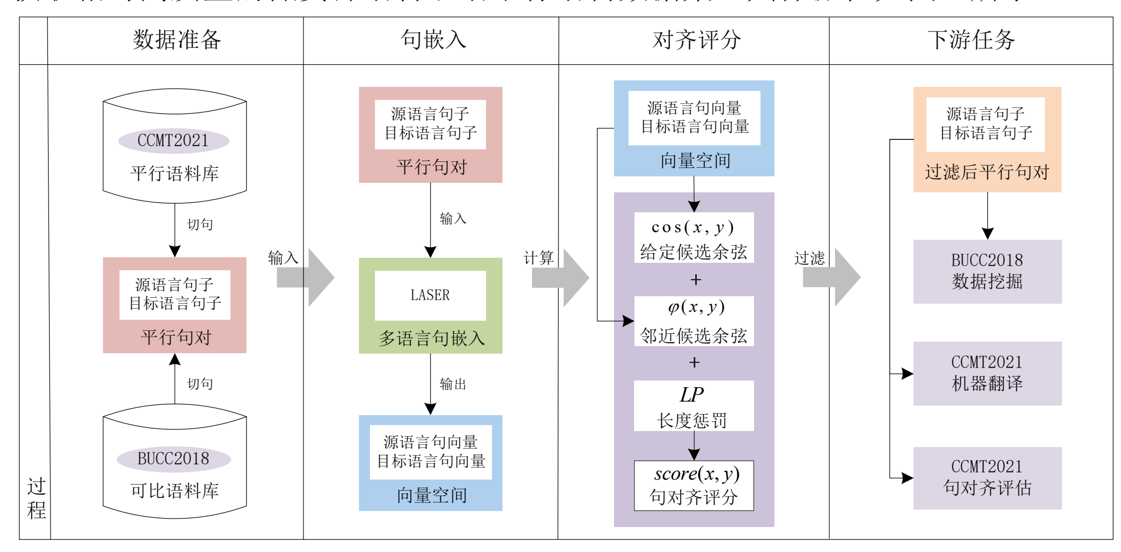


图 1　低资源平行语料句对齐流程图

Fig.1 Flowchart of Sentence Alignment for Low-resource Parallel Corpus

3.1　神经机器翻译

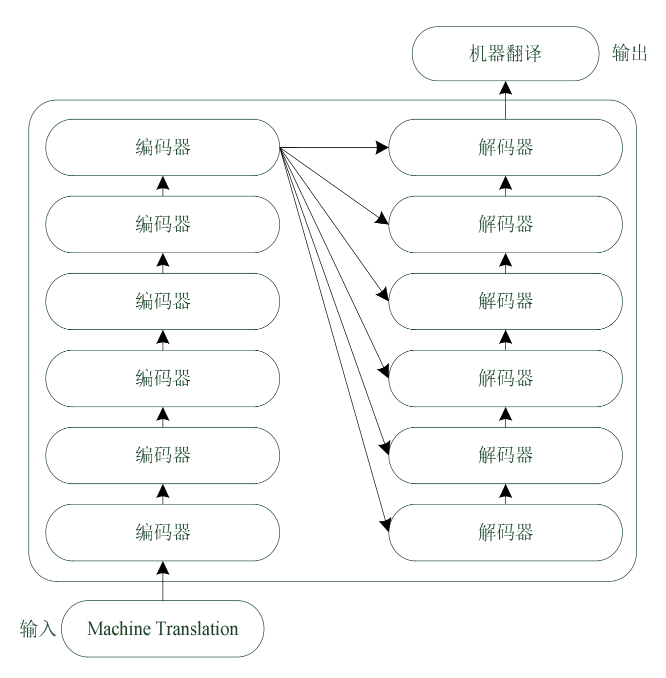


图 2　Transformer 模型架构

Fig.2 Architecture of Transformer Model

NMT 是一种序列到序列的生成问题，通过深度神经网络模型来实现源语言到目标语言的自动翻译。NMT 模型由编码器（Encoder）和解码器（Decoder）组成，其中，编码器将源语言句子 $x = (x_1, x_2, ..., x_n)$ 编码为一个连续向量，解码器将该

向量作为输入，生成目标语言的翻译结果 $y=(y_1,y_2,...,y_m)$。NMT 的模型训练使用了大规模的双语平行语料，通过最大化目标语言句子的概率优化模型参数 θ：

$$P(y|x)=\prod_{t=1}^{T}p(y_t|y_1,y_2,...,\ y_{t\text{-}1},\ \ x;\theta) \quad (1)$$

主流的神经网络模型包括循环神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Network，CNN）、基于注意力机制（Attention Mechanism）的 Transformer[26]模型等。本文采用了广泛使用的 Transformer 模型，该模型引入了自注意力机制（Self-attention）和多头自注意力机制（Multi-head Self-attention），用于更好捕捉句子的上下文信息。Transformer 同样遵循了编-解码架构，其中分别对编码器和解码器进行多层堆叠，如图 2 所示。

图 2 中的每一个编码器包含两个子层：多头自注意力层（Multi-head Self-attention）和前馈神经网络层（Feed Forward Neural Network），每一个子层都加入了残差连接（Residual Connections）和层归一化（Layer Normalization）。解码器除了与编码器相同的两个子层外，还新插入了一个编码器-解码器注意力层（Encoder-Decoder Attention）作用于编码器的输出。解码器同样在每个子层外使用了残差连接和层归一化。

3.2　多语言句嵌入

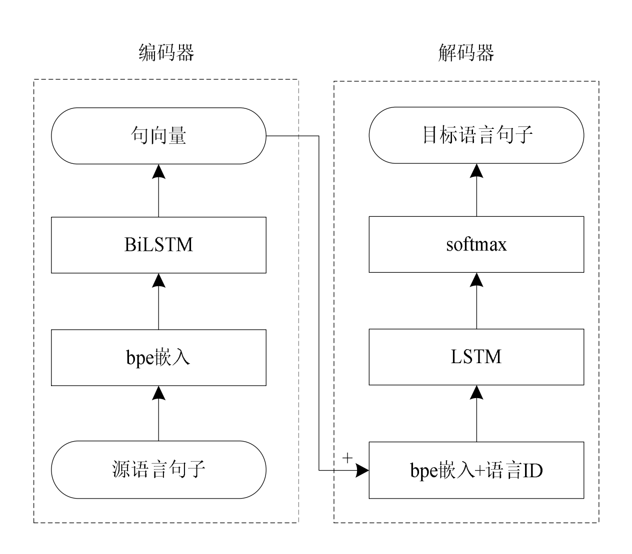


图 3　多语言句嵌入模型架构

Fig.3 Multilingual Sentence Embedding Model Architecture

多语言句嵌入将多种语言的完整句子嵌入到一个联合向量空间中，然后根据不同语言句子之间的语义距离来过滤或挖掘不同语言对之间的平行语料。本文的多语言句嵌入模型采用了开源工具 LASER[2]，该系统使用单个双向长短时记忆（Bidirectional Long Short-Term Memory, BiLSTM）编码器，该编码器不受语言限制，没有输入或输出语言的任何信号，并对所有语言共享了 40k 的字节对编码（Byte Pair Encoding, BPE）词表[27]；解码器在每个时间步骤都接受嵌入的输出语言 ID。编码器与辅助的解码器结合，在多个语言对上同时训练了一个序列到序列的系统，模型架构如图 3 所示。训练结束后，丢弃解码器，通过对所有编码器输出状态最大池化来获得定长向量的句子表示。该方法使用了固定维度，预训

2 https://github.com/facebookresearch/LASER

练阶段确定句子表示，在不反向传播到预训练模型的情况下对特定下游任务进行微调。这种方法在多种语言上进行了联合编码，具有跨语言一致性，可更好的运用于平行语料的挖掘。

3.3 对齐评分

（1）对齐评分

为了克服余弦相似度得分范围不一致的问题，Artetxe 和 Schwenk[6]提出了基于余弦相似度的差值评分方法，该方法主要考虑了给定候选句对的余弦相似度（给定候选余弦）与其最邻近的 k 个候选余弦相似度（邻近候选余弦）之间的比率差值。计算公式如下：

$$\text{score}(x,y)=\frac{\cos(x,y)}{\sum_{z\in \text{NN}_k(x)}\frac{\cos(x,z)}{2k}+\sum_{z\in \text{NN}_k(y)}\frac{\cos(z,y)}{2k}} \quad (2)$$

上式中，$\text{NN}_k(x)$ 表示源语言句向量 x 的 k 个邻近目标语言句向量 y（不包含重复句），$\text{NN}_k(y)$ 表示目标语言句向量 y 的 k 个邻近源语言句向量 x，一般设定 $k=4$。

$$\text{score}(x,y)=\frac{\cos(x,y)}{\varphi(x,y)}\times LP \quad (3)$$

其中：

$$\varphi(x,y)=\sum_{\bar{x}\in \text{NN}_k(x)}\frac{\cos(x,\bar{x})}{4k}+\sum_{u\in \text{NN}_k(x)}\frac{\cos(x,u)}{4k}+\sum_{v\in \text{NN}_k(y)}\frac{\cos(v,y)}{4k}+\sum_{\bar{y}\in \text{NN}_k(y)}\frac{\cos(y,\bar{y})}{4k} \quad (4)$$

公式 2 的差值越大表示句对齐评分越高，即两种语言的句子在语义上更加接近。由于在同一向量空间中邻近候选之间不总聚集，如图 4 中向量 y 的目标邻近候选余弦 $y2x$，因此，我们对公式 2 进行了改进，详见公式 3-4。

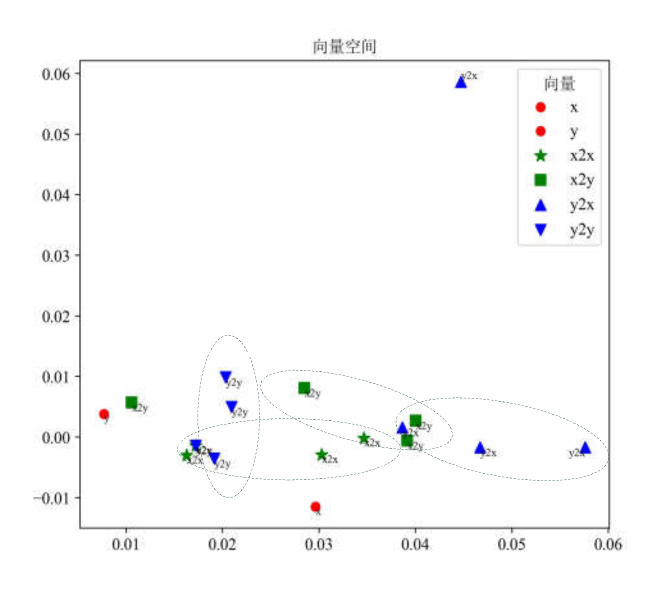


图 4　藏汉双语句嵌入向量示例

Fig.4 Example of Tibetan-Chinese Bilingual Sentence Embedding Vectors

为了在嵌入空间中拉近给定候选余弦与邻近候选余弦的差值，我们对源语言和目标语言分别进行了平滑修正，在邻近候选余弦中加入了给定源语言和目标语言句向量与自身邻近候选句向量（候选句向量中，不包含源语言和

目标语言句向量本身）之间的余弦相似度，详见公式 3 和 4，其中候选示例如图 5 所示。此外，公式 3 还对平行句对进行了句长惩罚（Length Penalty，LP）。



图 5　给定候选句对和邻近候选句对示例

Fig.5 Example of Given Candidate Sentence Pairs and Nearby Candidate Sentence Pairs

（2）句长惩罚

实验发现使用句对齐评分后，一些极短句的对齐评分相对较高，长句的对齐评分则相对靠后，导致通过评分排序过滤句对齐质量较差的句子时，长句更容易被过滤。我们认为数据集中过长或过短的句子比例失衡对机器翻译模型有一定干扰。

表 1　藏汉平行句对的对齐评分示例

Table1 Alignment Scores for Tibetan-Chinese Parallel Sentence Pairs

平行句对	不考虑句长	考虑句长
S:ལག་རྩལ་མཐོ་གསར་གྱི་ཐོན་ལས་གོང་འཕེལ་གཏོང་རྒྱུར་ཤུགས་རྒྱག་དགོས། T:要发展高新技术产业。	0.52	0.33
S:ནོར་སྲིད་གཏོང་སྒོ་དང་སྲིད་གཞུང་གི་མ་དངུལ་འཇོག་པའི་གྲུབ་ཚུལ་སྔ་མཐུད་ལེགས་སྒྲིག་བྱས། T:进一步优化了财政出入和政府投资体系。	0.86	0.77
S:རྒྱལ་ཁབ་ཀྱི་ཚོད་ལྟ་ཁང་དང་། རྒྱལ་ཁབ་ཀྱི་ལས་གྲྭའི་ལྟེ་གནས། ཁེ་ལས་བཅས་ལ་ཕྱོགས་པའི་གསར་གཏོད་འདེགས་སྐྱོར་ལས་སྟེགས་དང་། ཁེ་ལས་ཀྱི་ལག་རྩལ་ལྟེ་གནས་ཁག་ཅིག་གཙོ་གནད་དུ་བཟུང་སྟེ་འཛུགས་སྐྲུན་བྱེད། T:重点要建设一部分国家实验室、国家工程中心、面向企业的创新支撑平台和企业技术中心。	0.85	1.89

因此，为了消除句子长度的干扰，本文对所有给定平行句对进行了句长惩罚，通过惩罚使过短或过长的句子在计算对齐评分时不会占据过大的优势或劣势, 详见公式 5。

$$LP = \frac{l}{\bar{l}} (5)$$

上式中 LP 指给定平行句对的平均长度占整体平行语料库平均长度的比重。其中，分子 l 是给定候选平行句对的平均句长，分母 $\bar{l}$ 是语料库中所有平行句对的平均句长。表 1 给出了考虑长度因素后句对齐的评分示例，可以看出长度惩罚

降低了短句评分，提升了长句评分，进而使语料中短句和长句在进行句对齐评分时更加客观，不至于将更多的高分集中到贡献度较低的短句。需要注意的是后续实验部分我们只在机器翻译相关的任务中使用了该项。

（3）候选策略

在生成目标邻近候选句对时，我们采用了与 Artetxe 和 Schwenk[6]相同的四种策略。其中，

前向检索：每个源语言句子恰好与一个得分最高的目标句子对齐，有些目标句子可能与多个语言句对齐，也可能没有。

后向检索：与前向策略相同，但方向相反。

交叉检索：前向和后向的候选交集，舍弃对齐不一致的句子。

最大检索：前向和后向候选的组合，选择得分最高的候选句。

4　实验任务及结果分析

本文在 BUCC2018 平行语料挖掘[3]、CCMT2021 机器翻译和双语句对齐评分任务上分别进行了实验。

4.1　BUCC 平行语料挖掘

构建和使用可比语料库（Building and Using Comparable Corpora，BUCC）是为平行语料挖掘建立的评估任务，指给定两个不同语言的可比语料库，从中识别彼此翻译的句子对。任务设定挖掘英语到四种语言（德语、法语、俄语、汉语）之间的平行句对，每种语言包含 15 万-120 万句子，分为样本集、训练集和测试集，其中包含大约 2-3%的句子平行。

为了与 Artetxe 和 Schwenk[6]的结果对比，我们仅在英语-德语、英语-法语平行语料挖掘任务上进行了对比，表 2 展示了 NeuroAlign 方法在 BUCC2018[6]训练集上的准确率、召回率和 F1 值。

表 2　BUCC（2018）平行语料挖掘实验结果

Table2 BUCC (2018) Parallel Corpus Mining Experiment Results

检索方法	英语-德语						英语-法语					
	差值评分			NeuroAlign			差值评分			NeuroAlign		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
前向	95.2	94.4	94.8	**95.4**	**95.4**	**95.4**	92.4	91.3	91.8	92.0	**92.6**	**92.3**
后向	95.2	94.4	94.8	95.2	**95.8**	**95.5**	92.3	91.3	91.8	91.8	**93.0**	**92.4**
交叉	95.3	94.4	94.8	95.2	**95.9**	**95.5**	92.4	91.3	91.9	92.4	**92.6**	**92.5**
最大	95.3	94.4	94.8	95.2	**95.9**	**95.6**	92.4	91.3	91.9	92.0	**93.0**	**92.5**

实验结果表明，在英语-德语的平行语料挖掘任务中，NeuroAlign 方法准确率可保持在 95%以上，召回率可提升 1-1.5，F1 值可提升 0.6-0.8。在英语-法语平行语料挖掘中，NeuroAlign 方法准确率可保持在 92%左右，召回率可提升 1.3-1.7，F1 值可提升 0.5-0.6。相比之下，NeuroAlign 句对齐评分更有助于平行语料的挖掘和筛选，进一步验证了该方法的应用性能。

由于表 2 中“最大”检索方法的性能表现最佳，因此在后续实验中统一采用了该检索方法来生成目标邻近候选句对。

4.2　低资源语言机器翻译

3 https://comparable.limsi.fr/bucc2018/bucc2018-task.html

（1）实验数据和模型参数

该任务在 CCMT2021 双语翻译任务中的藏汉、维汉、蒙汉数据集上进行了机器翻译系统训练。其中，藏汉、维汉、蒙汉的验证集分别为 CCMT2017、CCMT2018、CCMT2017，数据规模详见表 3。其中，我们对藏语句子进行了分音节处理，对所有汉语句子进行了分字处理，并对所有语言对进行了 BPE 处理。

表 3　CCMT2021 训练数据集

Table3 CCMT2021 Training Dataset

语言对	原始数据	去重	过滤
藏汉	156578	148334	140918
维汉	170061	166062	157759
蒙汉	255824	250120	237614

关于实验设置，本文使用的神经机器翻译模型 Transformer 来源于 Facebook 的开源工具 Fairseq[28]。其中，我们使用了 6 层的编码器和解码器；选择了 Adam 优化器，betas 参数设置为（0.9，0.98）；学习率为 0.0005，dropout 为 0.3，批处理大小为 4096，所有模型的最大训练轮次均为 100；解码时 Beam size 设置为 4，其他参数采用了 Fairseq[4]中的默认设置。

（2）基线系统实验对比

我们在四个基线系统上进行了低资源机器翻译的实验对比，每个系统代表用不同的方法对低资源语言对（藏汉、蒙汉、维汉）进行句对齐评分，并根据评分排序，过滤评分相对较低的平行句对，然后在同一 NMT 模型架构上分别训练了翻译系统。其中：

Baseline：从去重语料中随机采样 95%的平行句对；

Cos：用余弦相似度进行句对齐评分；

Margin：用 Artetxe 和 Schwenk[6]的差值评分方法进行句对齐评分；

NeuroAlign：用公式 3 进行句对齐评分，此处不考虑句长惩罚。

为了保证低资源语言机器翻译的语料规模，本文中我们只过滤了评分后 5%对齐质量较差的平行句对，实验结果如表 4 所示。

表 4　不同对齐方法过滤语料的 BLEU 值

Table4 BLEU Scores of Filtered Corpora using Different Alignment Methods

方法	藏汉		维汉	蒙汉		
	CCMT2018	CCMT2019	CCMT2018	CCMT2017	CCMT2018	CCMT2019
Baseline	36.8	22.6	37.25	32.62	59.51	41.11
Cos	36.59	29.85	37.63	31.87	59.37	41.93
Margin	36.74	27.53	37.84	31.47	59.21	41.44
NeuroAlign	**37.11**	**33.47**	**37.91**	31.52	59.1	**41.18**

从表 4 可以看出，通过语料过滤的方式可以提升 NMT 的性能，说明语料的对齐质量对机器翻译至关重要。用本文提出的句对齐评分方法过滤评分较低的句子，系统的翻译性能提升效果明显，这一点在藏汉和维汉机器翻译中能够得以验证。其中：

相比于 Baseline，NeuroAlign 评分方法在藏汉翻译任务中的 BLEU 值分别提

4 https://github.com/facebookresearch/fairseq

升了 0.31、10.87，维汉翻译任务中提升了 0.66；蒙汉 CCMT2019 翻译任务中提升了 0.07。

与 Cos 相比，NeuroAlign 评分方法在藏汉翻译任务上分别提升了 0.52、3.62，维汉翻译任务上提升了 0.28。

与 Margin 相比，NeuroAlign 评分方法在藏汉翻译任务上提升了 0.37、5.94，维汉翻译任务上提升了 0.07，蒙汉 CCMT2017 翻译任务上提升了 0.05。总体来看，在该项任务中，我们提出的 NeuroAlign 评分可以有效过滤语料质量较差的平行句对，提升翻译性能。

但在蒙汉翻译任务中 Baseline 取得了较好的翻译效果，其他评分方法均有不同程度的翻译性能下降。为探究这一原因，我们对不同方法评分后的蒙汉平行语料对比发现，Baseline 的训练语料中，短句对随机均匀分布在语料中；而采用句对齐评分方法排序后的语料中，通常短句对评分较高，长句对评分较低。其中，句长在 1-5 的短句评分占据高分，从而不易被过滤，但是这些语料对翻译系统的语义贡献不够高，且对计算资源的利用不够充分。

表 5　蒙汉训练语料前 1 万句中短句对（1-5）的平均句长

Table5 Average Sentence Length of Short Sentence Pairs (1-5) in the First 10,000 Sentences of Mongolian-Chinese Training Corpus

方法	平均句长
Baseline	4.10
Cos	3.32
Margin	3.09
NeuroAlign	3.20
Margin+LP	4.25
NeuroAlign+LP	4.35

可以从表 5 看出，蒙汉训练语料的前 1 万个平行句对中，Baseline 短句对的平均句长高于其他系统的训练语料，而增加句长惩罚后，前 1 万个平行句对中，Margin+LP 和 NeuroAlign+LP 系统短句对的平均句长分别提升了 1.16、1.15。

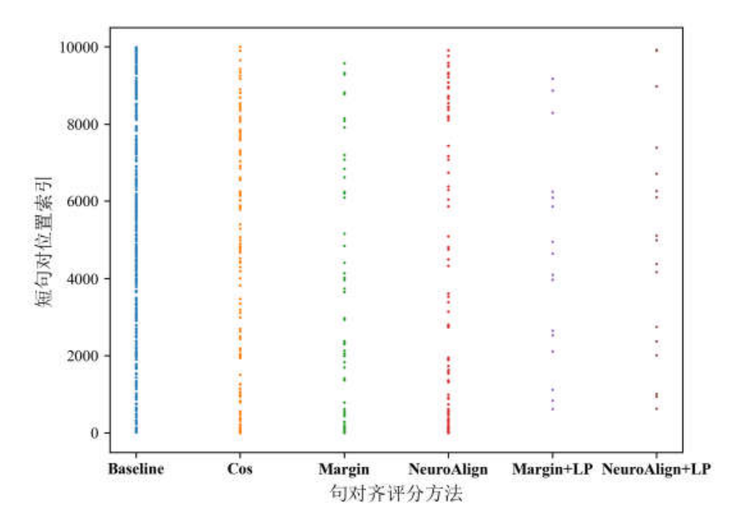


图 6　蒙汉训练语料前 1 万句中短句对（1-5）的位置分布

Fig.6 Distribution of Short Sentence Pairs (1-5) in the First 10,000 Sentences of the Mongolian-Chinese Training Corpus

同样，图 6 展示了蒙汉训练语料前 1 万个平行句对中短句对的位置索引，可

以看出，短句对在 Baseline 的训练语料中随机均匀分布，而用 Cos、Margin 对齐评分后短句对的分布相对靠前，故而无法按照评分排序对其进行过滤。因此，我们对句长进行了惩罚，意图保留更多语义贡献较高的句对，可以从图 6 中看出，增加长度惩罚后 Margin+LP 和 Remargin+LP 的短句明显减少了。

因此，有理由认为使用句对齐评分后，通常短句评分较高，长句评分较低，在语料过滤时会删除过多的长句，使得翻译系统对自然语言的语义捕获不够丰富，从而降低了系统的翻译性能。为了验证这一影响，我们在 Margin 和 NeuroAlign 评分中加入了句长惩罚，并进行了消融实验对比。

（3）消融实验对比

表 6 消融实验 BLEU 值对比

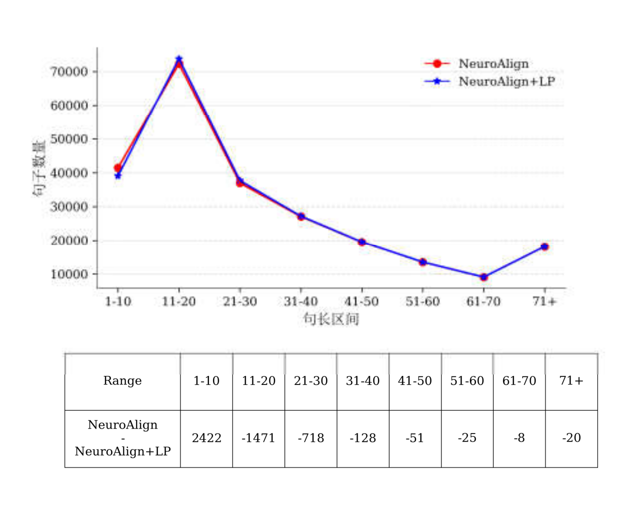
Table6 BLEU Scores Comparison in Ablation Experiments

方法	藏汉		维汉	蒙汉		
	CCMT2018	CCMT2019	CCMT2018	CCMT2017	CCMT2018	CCMT2019
margin	36.74	27.53	37.84	31.47	59.21	41.44
margin+LP	**37.24**	**30.19**	**38.48**	**32.39**	59.1	**41.74**
remargin	37.11	33.47	37.91	31.52	59.1	41.18
remargin+LP	**37.72**	**34**	**38.11**	**31.86**	**59.23**	**41.53**

从消融实验结果表 6 可以看出，与 Margin 和 NeuroAlign 的评分方法相比，增加句长惩罚后，藏汉、维汉、蒙汉的翻译效果均有了不同程度的提升。其中：

与 Margin 相比，Margin+LP 在藏汉上分别提升了 0.5、2.66，维汉提升了 0.64，蒙汉在 CCMT2017、CCMT2019 翻译任务上分别提升了 0.92、0.3。

与 NeuroAlign 相比，NeuroAlign+LP 在藏汉上分别提升了 0.61、0.53，维汉提升了 0.2，蒙汉翻译任务上分别提升了 0.34、0.13、0.35。上述结果表明，句长惩罚对语料的句对齐评分有着积极的影响，为进一步探究影响，我们对加入长度惩罚因子前后的语料句长分布进行了对比。



Range	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71+
NeuroAlign - NeuroAlign+LP	2422	-1471	-718	-128	-51	-25	-8	-20

图 7 蒙汉训练语料平均句长的分布

Fig.7 Distribution of Average Sentence Length Variation in Mongolian-Chinese Training Corpus

从图 7 可以看出，在 NeuroAlign 评分方法中加入长度惩罚后，蒙汉语料中句长为 1-10 的短句减少了 2422 个句对，而其他句长区间的句对则分别增加了 1471、718、128、51、25、8、20。因此，句长惩罚的效果明显，可以有效过滤短句，保留相应数量的长句，为翻译系统贡献更多语义完整的句子。

表 7　过滤语料示例

Table7 Examples of Filtered Corpus

语言	示例
藏汉	S1:འདི། …… T1:这个…… score：NeuroAlign:1.07、NeuroAlign+LP：0.58 S2:0.4%ཟིན། T2:占 0.4%， score：NeuroAlign:1.06、NeuroAlign+LP：0.29
维汉	S1: كىم كەلدى ؟ T1:谁来了？ score：NeuroAlign:1.18、NeuroAlign+LP：0.39 S2: سېنىڭ ئۆيۈڭدە T2:你家中。 score：NeuroAlign:1.14、NeuroAlign+LP：0.25
蒙汉	S1: [illegible]？ T1:啊？ score：NeuroAlign:1.06、NeuroAlign+LP：0.31 S2:[illegible]! T2:精彩！ score：NeuroAlign:0.91、NeuroAlign+LP：0.25

表 7 中展示了过滤的短句示例，这些短句在 NeuroAlign 评分中占据高分，属于对齐质量较好的句子，不易被过滤；但在 NeuroAlign+LP 评分中排分则相对靠后，由于这些句子携带的语义不够丰富，我们认为可以被过滤。

表 8　不同消融系统的译文对比

Table7 Translations Comparison of Different Ablation Systems

语言	译文对比
藏汉	S：དོན་འབྲས་ལྡན་པའི་མཉམ་ལས་སྤེལ་རྒྱུ་ནི་ཧྲང་ཧའི་མཉམ་ལས་ཚ་འཛུགས་གོང་འཕེལ་ཡོང་བའི་དངོས་པོའི་རྨང་གཞི་དང་ཐོག་མའི་སྒུལ་ཤུགས་རེད། T:**务实合作**是上海合作组织发展的物质基础和**原动力。** 译文 1：开展**有效合作**是上海合作组织发展的物质基础和**初次动力**。（NeuroAlign） 译文 2：开展**有成果合作**是上海合作组织发展的物质基础和**首动力**。（NeuroAlign+LP）
维汉	S：ئەنگلىيەدە بۇ يىل كۆپ قېتىم تېرورلۇق ۋەقەلىرى يۈز بېرىپ، ئېغىر ئۆلۈم-يېتىم بولدى. T:英国今年发生了多起**恐怖袭击**事件，造成严重人员伤亡。 译文 1：英国今年多次发生**暴力**事件，造成严重人员伤亡。（NeuroAlign） 译文 2：英国今年发生多起**恐袭**事件，造成重大人员伤亡。（NeuroAlign+LP）
蒙汉	S：[illegible] T:这就是我们从小汽车换乘公共汽车的地方。 译文 1：这是我们**的**小汽车换乘公共汽车的地方。（NeuroAlign） 译文 2：这是我们**从**小汽车换乘公共汽车的地方。（NeuroAlign+LP）

此外，可以从表 8 的消融实验示例看出，增加句长惩罚后，NeuroAlign+LP 系统对源语言的翻译用词更接近参考译文，如藏汉翻译中，将“原动力”翻译为“首动力”，将“务实合作”翻译成“有成果合作”；在维汉翻译中将“恐怖袭击”翻译为“恐袭”。此外，对于语义的捕获也更加准确，如：在蒙汉翻译中对“从”

字句翻译准确，而 Margin 系统则是捕获成了领属关系。上述实验进一步证实了句长惩罚因子对句对齐评分的积极贡献。

4.3 低资源语言句对齐评分

该任务对低资源语言 NMT 的训练语料进行了句对齐评分。表 9 以“最大”检索策略为例，将 CCMT2021 中藏汉、蒙汉、维汉所有语言对上的句对齐评分与人工评分进行了对比。

表 9　CCMT2021 训练语料对齐评分

Table9 Alignment Scores of CCMT2021 Training Corpus

方法	藏汉	维汉	蒙汉
Cos	50.8	39.9	45.8
Margin	76.9	57.5	64.4
Margin+LP	77.9	56.9	69.6
NeuroAlign	67.7	52.5	56.8
NeuroAlign+LP	68.9	51.9	61.7

可以看出，Cos 句对齐评分整体较低，Margin 差值评分时所有数据集的评分整体上升，而用本文提出的方法评分后分值又呈现出下降趋势，说明我们提出的方法更接近于 Margin 差值评分方法，但评分更加严格。

表 10　不同方法在藏汉测试句上的对齐评分

Table10 Alignment Scores of Different Methods on Tibetan-Chinese Test Sentences

方法	评分
Cos	50.5
Margin	74.1
NeuroAlign	64.8
NeuroAlign +LP	81.0
Human	79.98

我们从 CCMT2021 藏汉数据集中人工抽取了 100 个测试句对，包含未对齐、未翻译、语言错误、翻译错误、断句错误、编码错误等六类未对齐现象，每个句对至少包含其中的一类错误。我们请母语人对这些测试句对进行标记评分，每个句对总计 6 分，出现一类错误扣除 1 分，然后以总得分的百分比换算人工评分。为了保证人工评分的客观性，我们请 2 名藏语母语人同时进行了评分，取平均值为最终人工评分。表 10 展示了平行测试句对的自动评分和人工评分，对比结果发现，我们提出的评分方法更接近人工评分。

5 结语

本文提出了一种基于神经网络的无监督句嵌入双语平行语料句对齐评分方法，在平行语料挖掘、低资源 NMT 和句对齐评分等自然语言处理任务上实验结果表明，我们提出的评分方法可对低资源双语平行语料的句对齐质量进行有效评分，且根据评分排序过滤对齐质量较差的语料后，可以有效提升 NMT 系统的翻译性能，该方法在高资源的平行语料挖掘中同样适用。但是限于低资源双语平行语料的资源匮乏，我们未在藏汉、维汉、蒙汉以外的语言对上进行更多探索。此外，为进一步考虑平行语料数据质量对机器翻译系统的影响，我们认为句对齐只是平行语料质量中的重要指标之一，其他质量因素对机器翻译也存在影响。未来

我们将在更多丰富的语言对上进行探索，同时也会对面向机器翻译的语料质量评估做出更多的指标探索，以期对机器翻译的语料质量评估起到重要推动作用。

参考文献：

[1] Koehn P, Knowles R. Six Challenges for Neural Machine Translation[C]//Proceedings of the First Workshop on Neural Machine Translation. 2017: 28-39.

[2] Goutte C, Carpuat M, Foster G. The Impact of Sentence Alignment Errors on Phrase-based Machine Translation Performance[C]//Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers. 2012.

[3] Khayrallah H, Koehn P. On the Impact of Various Types of Noise on Neural Machine Translation[C]//2nd Workshop on Neural Machine Translation and Generation. Association for Computational Linguistics, 2018: 74-83.

[4] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[5] Liu Y, Gu J, Goyal N, et al. Multilingual Denoising Pre-training for Neural Machine Translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[6] Artetxe M, Schwenk H. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3197-3203.

[7] Artetxe M, Schwenk H. Massively Multilingual Sentence Embeddings for Zero-shot Cross-lingual Transfer and Beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610.

[8] Schwenk H. Filtering and Mining Parallel Data in a Joint Multilingual Space[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 228-234.

[9] Shi L, Niu C, Zhou M, et al. A Dom Tree Alignment Model for Mining Parallel Data from the Web[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006:489-496.

[10] Ma S, Zhang C. Automatic Collection of the Parallel Corpus with Little Prior Knowledge[C]//International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer International Publishing, 2014: 95-106.

[11] Richard Bellman. An Introduction to the Theory of Dynamic Programming[J]. Technical Report, RAND Corporation, Santa Monica, CA. 1953: 0104.

[12] Brown P F, Lai J C, Mercer R L. Aligning Sentences in Parallel Corpora[C]//29th Annual Meeting of the Association for Computational Linguistics. 1991: 169-176.

[13] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corpora[J]. Computational Linguistics, 1994, 19(1): 75-102.

[14] Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora[C]//Conference of the Association for Machine Translation in the Americas. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 135-144.

[15] Varga D, Halácsy P, Kornai A, et al. Parallel Corpora for Medium Density Languages[J]. Amsterdam Studies In the Theory and History of Linguistic Science Series 4, 2007, 292: 247.

[16] Simard M, Foster G F, Isabelle P. Using Cognates to Align Sentences in Bilingual Corpora[C]//Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. 1992.

[17] Sennrich R, Volk M. MT-based Sentence Alignment for OCR-generated Parallel Texts[C]//In 9th Conference of the Association for Machine Translation in the Americas. 2010.

[18] Sennrich R, Volk M. Iterative, MT-based Sentence Alignment of Parallel Texts[C]//Proceedings of the 18th

Nordic conference of computational linguistics. 2011: 175-182.
[19] Schwenk H, Chaudhary V, Sun S, et al. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 1351-1361.
[20] Guo M, Shen Q, Yang Y, et al. Effective Parallel Corpus Mining using Bilingual Sentence Embeddings[C]//Proceedings of the Third Conference on Machine Translation: Research Papers. 2018: 165-176.
[21] Hangya V, Braune F, Kalasouskaya Y, et al. Unsupervised Parallel Sentence Extraction from Comparable Corpora[C]//Proceedings of the 15th International Conference on Spoken Language Translation. 2018: 7-13.
[22] Thompson B, Koehn P. Vecalign: Improved Sentence Alignment in Linear Time and Space[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 1342-1348.
[23] Zhang W. Improve Sentence Alignment by Divide-and-conquer[J]. arXiv preprint arXiv:2201.06907, 2022.
[24] Johnson M, Schuster M, Le Q V, et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-shot Translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.
[25] Papineni K, Roukos S, Ward T, et al. Bleu: A Method for Automatic Evaluation of Machine Translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[J]. Advances in neural information processing systems, 2017, 30.
[27] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]//54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2016: 1715-1725.
[28] Ott M, Edunov S, Baevski A, et al. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling[C]// Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

通讯作者（Corresponding author）：赵小兵（Zhao Xiaobing），
E-mail:nmzxb_cn@163.com。

基金项目：本文系“国家社会科学”基金项目（项目编号：22&ZD035）的研究成果之一。
The work is supported by The National Social Science Fund of China (Grant No. 22&ZD035).

作者贡献声明：

李林霞：设计研究方案，实验,论文撰写；

陈波：提出研究思路，修改论文；

周毛克：语料核对，数据分析；

赵小兵：论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1]构建和使用可比语料库数据集（BUCC2018）.

https://comparable.limsi.fr/bucc2018/bucc2018-task.html

[2]李林霞. 低资源语言平行语料句对齐评分数据集.

DOI:10.57760/sciencedb.j00133.00298.